Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp 141-144 recibido 30-03-2017 revisado 11-05-2017 aceptado 29-05-2017

Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje

Efforts to foster biomedical text mining efforts beyond English: the Spanish national strategic plan for language technologies

Marta Villegas[1], Santiago de la Peña[2], Ander Intxaurrondo[2], Jesus Santamaria[2], Martin Krallinger[2*]
[1]Barcelona Supercomputing Center (BSC). Jordi Girona, 29 08034 Barcelona
[2]Centro Nacional de Investigaciones Oncológicas (CNIO)
Melchor Fernández Almagro, 3 28029 Madrid
marta.villegas@bsc.es
{sdelapena,aintxaurron,jsantamaria,mkrallinger}@cnio.es

Resumen: Si bien se han hecho esfuerzos considerables para aplicar las tecnologías de minería de texto a la literatura biomédica y los registros clínicos escritos en inglés, lo cierto es que intentos de procesar documentos en otros idiomas han atraído mucha menos atención a pesar de su interés práctico. Debido al considerable número de documentos biomédicos escritos en español, existe una necesidad apremiante de poder acceder a los recursos de minería de textos biomédicos y clínicos desarrollados para esta lengua de alto impacto. Para abordar este asunto, la Secretaría de Estado encargó las actuaciones de apoyo técnico especializado para el desarrollo del Plan de Impulso de las tecnologías del Lenguaje en el ámbito de la biomedicina. El artículo describe brevemente las líneas principales de actuación del proyecto en su primera fase, esto es: facilitar el acceso a recursos y herramientas en PNL, analizar y garantizar la interoperabilidad del sistema, la definición de métodos y herramientas de evaluación, la difusión del proyecto y sus resultados y la alineación y colaboración con otros proyectos nacionales e internacionales. Además, hemos identificado algunas de las tareas críticas en el procesamiento de textos biomédicos que requieren investigación adicional y disponibilidad de herramientas.
Palabras clave: Text mining, minería de textos, plan de impulso, infraestructuras lingüísticas, recursos lingüísticos.

Abstract: A considerable effort has been made to apply text mining technologies to biomedical literature and clinical records written in English, while attempts to process documents in other languages have attracted far less attention despite the key practical relevance. Due to the considerable number of biomedical documents written in Spanish, there is a pressing need to be able to access biomedical and clinical text mining resources developed for this high impact language. To address this issue, the Spanish Ministry of State for Telecommunications launched the Plan for Promotion of Language Technologies in the field of biomedicine with the aim of providing specialized technical support to research and development of software solutions adapted to this domain. This article briefly describes the main lines of action of this project in its initial stages, namely: (a) identification of relevant biomedical NLP resources/tools, (b) examining and enabling system interoperability aspects, (c) to outline strategies and support for evaluation settings, (d) to disseminate the project and its results, and (e) to align and collaborate with other related national and international projects. Moreover we have identified some of the critical biomedical text processing tasks that require additional research and availability of tools.
Keywords: Plan for promotion of language technologies, text mining, linguistic infraestructures, biomedical documents, clinical records.

ISSN 1135-5948 © 2017 Sociedad Española para el Procesamiento del Lenguaje Natural

1 Introducción y antecedentes

Las técnicas de minería de textos en literatura biomédica escrita en inglés han experimentado resultados significativos mientras que los intentos de procesar documentos en otros idiomas han atraído mucha menos atención a pesar de su interés práctico. Sin embargo, el considerable número de documentos biomédicos escritos en español, genera la necesidad apremiante de poder acceder a los recursos de minería de textos biomédicos y clínicos desarrollados también para esta lengua. Para abordar este asunto, la Secretaría de Estado encargó las actuaciones de apoyo técnico especializado para el desarrollo del Plan de Impulso de las Tecnologías del Lenguaje en el ámbito de la biomedicina.

Así pues, el proyecto que anunciamos se inscribe dentro del Plan de Impulso de las Tecnologías del Lenguaje de la Agenda Digital para España[1], aprobada en febrero de 2013 como la estrategia del Gobierno para desarrollar la economía y la sociedad digital. Esta estrategia se configuró como el paraguas de todas las acciones del Gobierno en materia de Telecomunicaciones y de Sociedad de la Información y marca la hoja de ruta en materia de Tecnologías de la Información y las Comunicaciones (TIC) y de Administración Electrónica para el cumplimiento de los objetivos de la Agenda Digital para Europa[2].

Para la puesta en marcha y ejecución de la Agenda se definieron diferentes planes específicos entre los que se encuentra el Plan de Impulso de las Tecnologías del lenguaje[3] que tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española y lenguas co-oficiales. Para ello, el Plan define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas co-oficiales.
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria.
- Incorporen a la Administración como impulsor del sector de procesamiento de lenguaje natural.

Así pues, el proyecto que describimos forma parte de la encomienda que la Secretaría de Estado encargó para la realización de las actuaciones de apoyo técnico especializado para el desarrollo del Plan en el ámbito de la biomedicina. En breve se habilitará el sitio web del proyecto y se anunciará en la web de la agenda digital.

2 Tareas

Los objetivos del proyecto incluyen los siguientes aspectos, con un enfoque especial al ámbito del procesamiento de documentos biomédicos/clínicos:

- La definición y fomento de estándares de interoperabilidad y de modelos de licencias.
- La especificación de requisitos para la protección de datos personales.
- El fomento y metodología para la reutilización de recursos.
- La supervisión y soporte a los diferentes proyectos de PLN (procesamiento del lenguaje natural) en biomedicina que surjan para garantizar que éstos se alinean con los objetivos del Plan.
- La creación de métodos y campañas de evaluación que potencien el desarrollo de infraestructuras lingüísticas biomédicas.

3 Líneas de actuación

En una primera fase, el proyecto gira entorno a cinco líneas básicas de actuación: facilitar el acceso a recursos y herramientas, garantizar la interoperabilidad del sistema, establecer métodos de evaluación y divulgar el proyecto. Además, se buscará establecer sinergias y colaboraciones con otros proyectos nacionales e internacionales con el fin de lograr el máximo impacto.

En adelante se describen brevemente las acciones a realizar durante este año para cada una de las líneas de trabajo.

3.1 Compilación de corpus biomédico

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con

[1] http://www.agendadigital.gob.es

[2] https://ec.europa.eu/digital-single-market/

[3] http://www.agendadigital.gob.es/tecnologias-lenguaje/Paginas/plan-impulso-tecnologias-lenguaje.aspx

licencia abierta que permita: ejecutar tareas de PLN sobre big data y replicar los experimentos. Para ello se contemplan diferentes acciones:

Creación de un agregador de publicaciones de acceso abierto en biomedicina. El proyecto partirá de la tarea realizada por otras iniciativas en el ámbito de las publicaciones científicas como son el buscador de ciencia abierta Recolecta[4], IBECS[5], MEDES[6], o Scielo[7], biblioteca virtual formada por una colección de revistas científicas españolas de ciencias de la salud. El objetivo es colaborar con estos buscadores para poder ir un paso más allá y convertir los diferentes repositorios digitales que éstos recolectan y agrupan en sus portales en un gran corpus biomédico. El sistema deberá poder indexar los artículos y permitir la creación de sub-corpus a demanda.

Se explorarán otras vías de agregación de contenidos textuales en biomedicina como la creación de un corpus de patentes, un corpus de informes médicos y otro de información farmacéutica. En este caso, el proyecto incentivará convenios de colaboración con organismos del sistema público sanitario y facilitará servicios de anonimización de datos para cumplir con los requisitos de la ley de protección de datos.

3.2 Recursos lingüísticos

El proyecto creará y mantendrá un catálogo estructurado de recursos específicos creados dentro del plan (recursos in house), como diccionarios léxico-semánticos, terminologías y listados de entidades de relevancia biomédica, tanto para el indexado de documentos como para diferentes modalidades y las técnicas de Extracción de Información. Se identificarán e incluirán también aquellos recursos externos que por su relevancia deban formar parte del catálogo de recursos del ámbito biomédico(Primo-Peña, 2016). El catálogo será compatible con el modelo de metadatos de META-SHARE[8] y con los catálogos de recursos de otros proyectos europeos como OpenMinTeD, CLARIN[9] y OLAC[10]. Para ello se generarán descripciones de metadatos en los diferentes esquemas cuando ello sea necesario.

3.3 Herramientas lingüísticas

El proyecto debe facilitar el uso e integración de herramientas de procesamiento de lenguaje natural y minería de textos. Se implementará un registro de servicios que permita la ejecución de los mismos. Para ello se identificarán las herramientas básicas que deben formar parte de cualquier aplicación de PLN, incluyendo herramientas de pre-proceso y herramientas lingüísticas.

Se evaluarán específicamente herramientas de minería de textos en biomedicina como MetaMap[11] (desarrollado por la Biblioteca Nacional de Medicina de EEUU), cTakes[12] (herramienta similar a Metamap desarrollada por Apache), i2b2[13] (desarrollada por el centro i2b2 y utilizada para detectar terminología médica y abreviaturas) o MedTagger[14] (parte de la OHNLP[15]). Todas las herramientas identificadas se describirán y incluirán en un registro disponible para la comunidad científica y la industria. En este contexto se llevará a cabo un estudio de interoperabilidad entre las herramientas del registro que permita definir las acciones a realizar para garantizar su correcta integración y compatibilidad. Se prestará especial atención a iniciativas similares con el fin de asegurar la máxima compatibilidad con otros proyectos y/o propuestas.

3.4 Evaluación

El proyecto dedicará especial atención a la evaluación, para ello se organizarán campañas de evaluación comparativa de herramientas de PLN (por ejemplo en el contexto de la competición de BioCreative[16] y IberEval[17]). Estas campañas potenciarán el desarrollo de infraestructuras lingüísticas en el área de la biomedicina de utilidad para el Plan y tendrán como resultado la creación de corpus Gold Standard reutilizables para la validación y el desarrollo de componentes de procesamiento del lenguaje natural en biomedicina, así como la definición de métricas

[4] https://www.recolecta.fecyt.es/#
[5] http://ibecs.isciii.es/
[6] https://www.medes.com/
[7] http://scielo.isciii.es/
[8] http://www.meta-net.eu/meta-share
[9] https://vlo.clarin.eu/?2
[10] http://www.language-archives.org/
[11] https://metamap.nlm.nih.gov/
[12] http://ctakes.apache.org/
[13] https://www.i2b2.org/index.html
[14] http://ohnlp.org/index.php/MedTagger
[15] http://www.ohnlp.org/index.php/Main_Page
[16] http://www.biocreative.org/
[17] http://sepln2017.um.es/ibereval.html

comparativas de validación. La infraestructura de evaluación será testeada en el contexto de campañas de evaluación y tiene como objetivo facilitar una validación de componentes con métricas estándar, así como ofrecer la posibilidad de visualizar anotaciones automáticas / manuales y proporcionar la generación de un informe de análisis de errores.

3.5 Interoperabilidad

El proyecto elaborará las recomendaciones y acciones necesarias para garantizar la interoperabilidad necesaria entre los distintos recursos y herramientas del sistema y así garantizar la reutilización y mantenimiento de infraestructuras lingüísticas en el área de la biomedicina. Se pondrá especial énfasis en asegurar el cumplimiento y desarrollo de estándares y especificaciones de interoperabilidad y compatibilidad para la integración de los recursos generados tanto de datos estructurados (recursos lingüísticos) como no-estructurados (corpus) de relevancia para el sector.

Para facilitar la interoperabilidad entre los diferentes recursos y entre éstos y las herramientas disponibles, se crearán los conversores de formato necesarios y se definirán las interfaces comunes de ejecución para las diferentes herramientas.

Se prestará especial atención a promover y garantizar la interoperabilidad con recursos y herramientas de otros proyectos del Plan.

3.6 Difusión

La difusión de los resultados del proyecto es clave para el fomento y el desarrollo de las tecnologías del lenguaje en este ámbito. Se prestará especial atención a la creación de tutoriales y manuales de buenas prácticas que avancen en el uso de estándares y métodos que garanticen la interoperabilidad de los futuros recursos del sistema. Con el fin de fomentar el uso del PLN se crearán calls y hackathons que sirvan de incentivo y ejemplo de uso.

4 Alineación con otros proyectos

Parte fundamental del proyecto es su alineación con proyectos nacionales (como la red ReTeLe[18]) e internacionales de relevancia en el ámbito. Así, se ha establecido ya colaboración con OpenMinTeD[19] y ELIXIR[20]. OpenMinTeD se propone crear una infraestructura abierta y orientada a servicios para la minería de texto y datos de contenido científico y académico. ELIXIR, por su parte, tiene por objetivo coordinar, integrar y mantener recursos en el ámbito de la bioinformática para su uso en la investigación.

El proyecto presta también especial atención a las actividades de la Research Data Alliance.

Bibliografía

Primo-Peña, E. 2016. Las bases de datos de información biomédica, ¿en español?: Presente y futuro. *Educación Médica*, 17(2):39–44.

Przybyłla, P., M. Shardlow, S. Aubin, R. Bossy, R. Eckart de Castilho, S. Piperidis, J. McNaught, y S. Ananiadou. 2016. Text mining resources for the life sciences. *Database*, 2016(0):baw145.

Rehm, G., J. Hajic, J. van Genabith, y A. Vasiļjevs. 2016. Fostering the next generation of european language technology: Recent developments - emerging initiatives - challenges and opportunities. En N. C. C. Chair) K. Choukri T. Declerck S. Goggi M. Grobelnik B. Maegaard J. Mariani H. Mazo A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Sarma, G. P. 2016. Scientific data science and the case for open access. *CoRR*, abs/1611.00097.

[18] http://retele.linkeddata.es

[19] http://openminted.eu/

[20] https://www.elixir-europe.org/